HDF5

Proposal for standardization

ESDSWG meeting
October 25 – 27, 2005
Baltimore, MD

Outline

- What is HDF5?
- Motivation and benefits for the HDF5 standard
- Overview of RFC
- Overview of HDF5 technology
- HDF5 existing implementations and stability

What is HDF5?

What is HDF5?

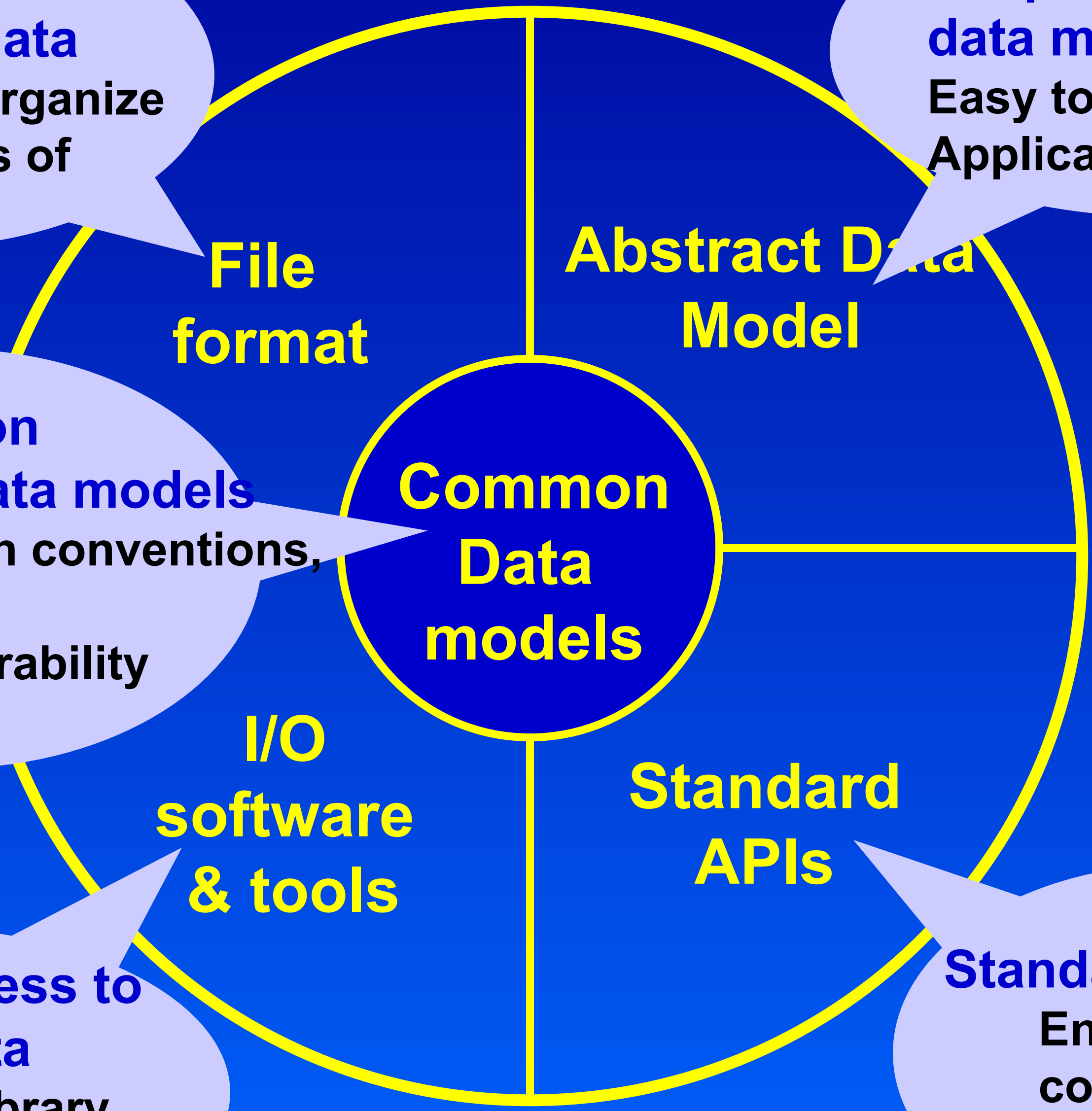
- Format for storing scientific data
  - To store and organize all kinds of data
- Software for accessing scientific data
  - Standard APIs, I/O library & tools
- Open source software
  - FreeBSD type of license
- Developed and supported by HDF group NCSA U of I
  - NASA ESDIS and ASCI DOE programs

Why HDF5?

Needed: a format to support data management in high end computing environments

- HDF4 (first release in 1989)
  - Rigid data model: array, image, palette, table, annotation, group
  - Limited to 2GB
  - Limited number of objects
- Flexible data model
- Better engineered and file format library
- Size and complexity of data
  - Big data sets and getting bigger
  - Variety of data types and structures
  - Metadata may be complex and big as data itself
- Efficient I/O (including parallel) and flexible storage

HDF5 solution



Common Data models

File format

Abstract Data Model

I/O software & tools

Standard APIs

Format for scientific data

Store, organize all kinds of data

Simple & powerful data model

Easy to represent Application data

Emphasis on standard data models

Establish conventions, improve interoperability

Efficient access to scientific data

Free I/O library, free & commercial tools

Standard Interfaces

Enforce conventions, avoid duplicate effort

Motivation for the HDF5 standard

Motivation for HDF5 standard

- Successful project supported by NASA ESDIS program
- HDF5 is underlying format for HDF-EOS 5 (Aura) and a distribution format for NPOESS
  - used for global climate change research
- Need to sustain HDF5 technologies to access HDF-EOS and NPOESS data in 10 and more years from now
- No official standards for general scientific binary data formats exists

Benefits of HDF5 standard

- HDF5 is widely used and is proven to be useful
  - 300 projects worldwide including NASA projects outside EOS (CGNS, PDEs JPL)
- HDF5 standardization will accelerate HDF5 adoption among EOS and NPOESS communities
- HDF5 standardization will validate HDF5 to vendors, government agencies and other organization

Benefits of HDF5 standard

- Increase adoption of HDF5 by government, academia and industry
- Improve usability of HDF5
  - More software for working with data in HDF5
  - More useful data stored in HDF5
- Leads to ANSI and ISO standards
- Extend across a broad range of science and engineering domains
- Commercial tools
  - IDL
  - Matlab
  - Mathematica
  - LabView

Earth Observing System Data & Information System (EOSDIS)

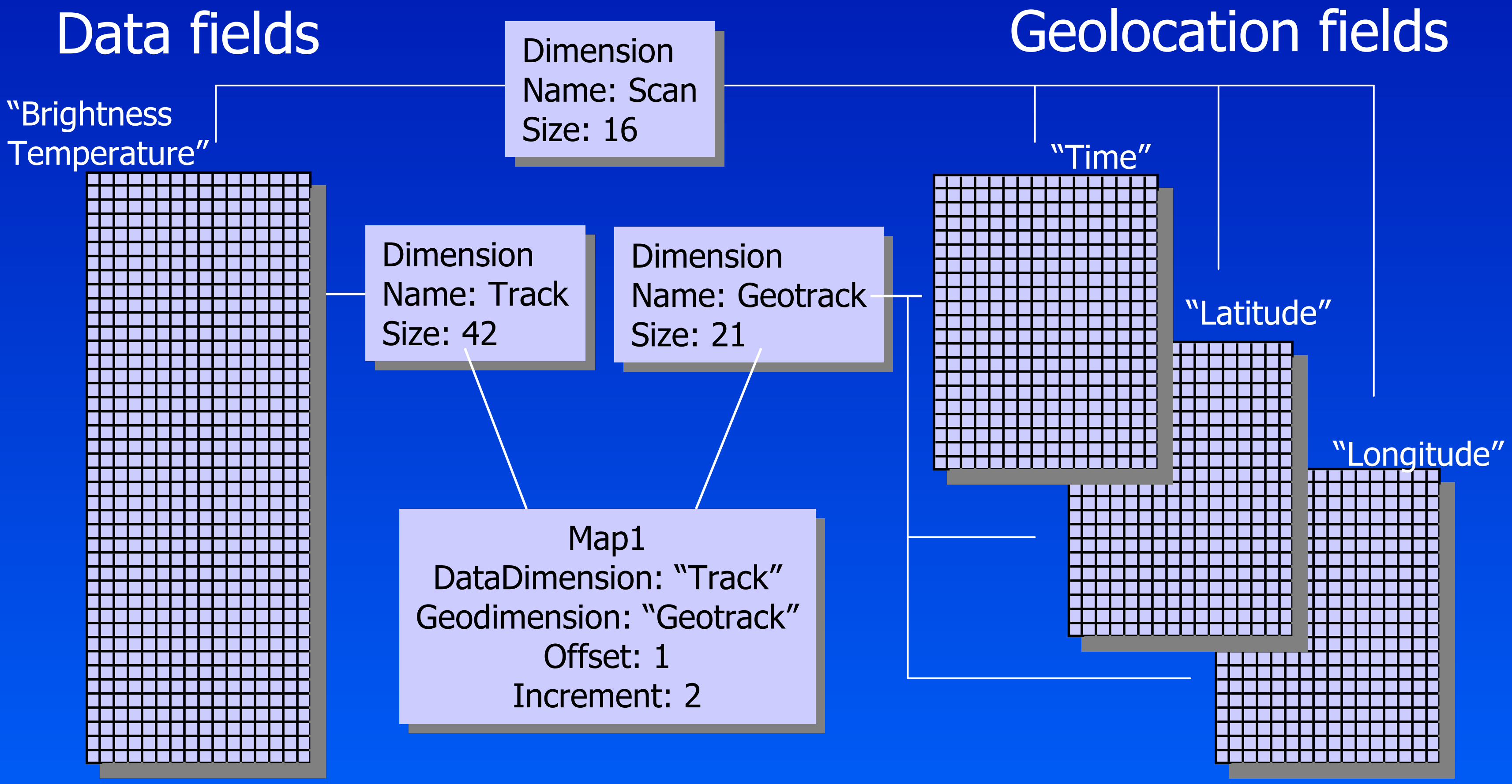
HDF-EOS example
HDF5 Standardization

- To share files, users must organize them similarly.
- HDF user groups create standard profiles
  - Ways to organize data in HDF files.
  - Metadata
  - API and library
- HDF-EOS example
  - Swath
  - Grid
  - Point

HDF_EOS Example
Standard profile API

- Hide the details of the underlying format
- Enforce standardization in use of format
- Provide data model that application understands
  - Data abstractions that make sense to application
  - Operations that application needs
- Common tools
- Data sharing
- Tune lower layers without changing application
- May even change the underlying format

HDF-EOS "Swath" profile



Data fields

Geolocation fields

Dimension
Name: Scan
Size: 16

"Brightness Temperature"

Dimension
Name: Track
Size: 42

Dimension
Name: Geotrack
Size: 21

"Time"

"Latitude"

"Longitude"

Map1
DataDimension: "Track"
Geodimension: "Geotrack"
Offset: 1
Increment: 2

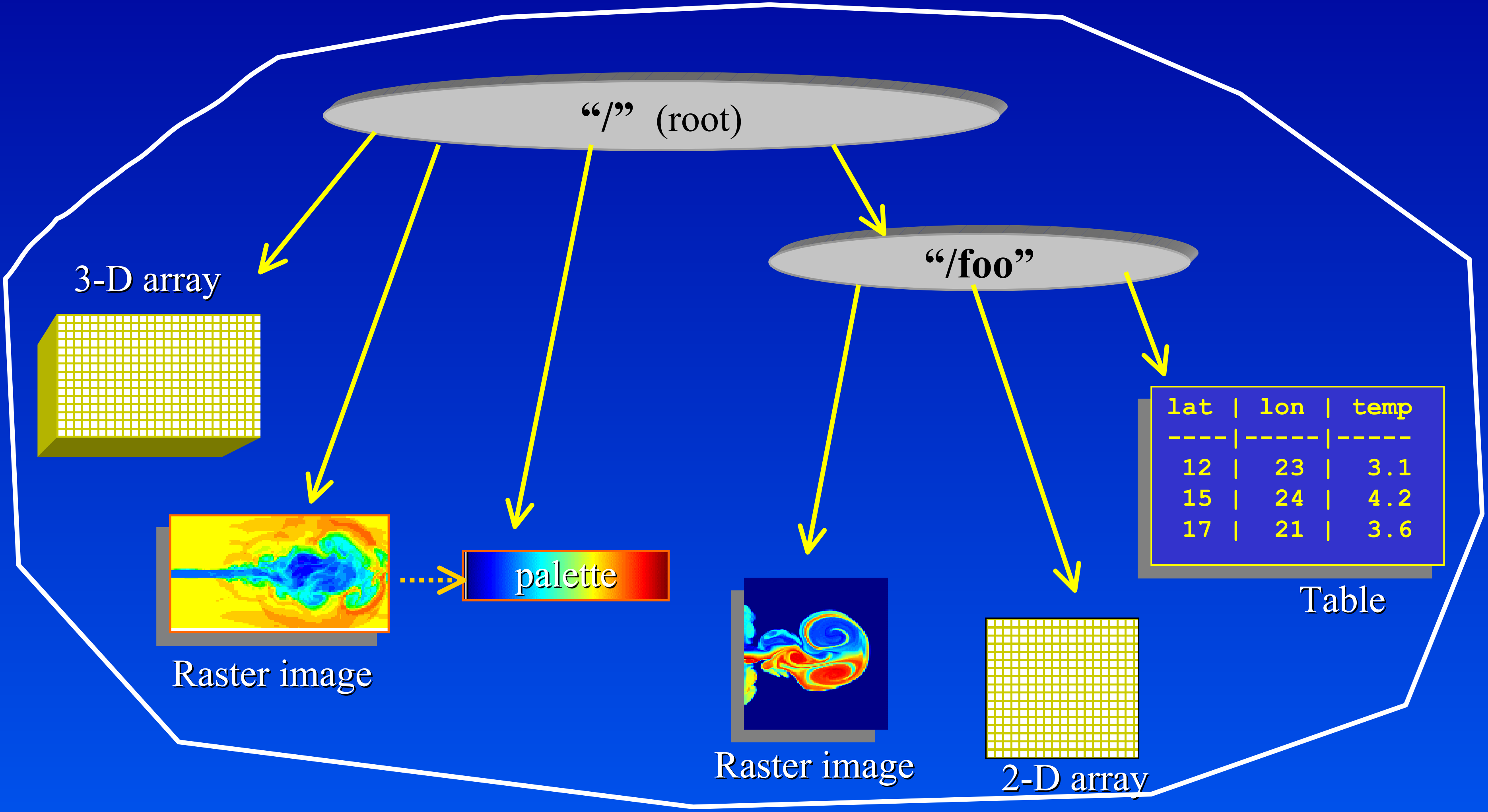
Overview of RFC

Overview of RFC

- Introduction to Technology and motivation for the standard
- HDF5 Data Model
  - Complete description
- HDF5 File Format
  - Overview
  - Points to the HDF5 File Format Specification
- HDF5 I/O Library
  - Overview
  - Points to the HDF5 Reference Manual for library APIs

HDF5 Data Model

Example HDF5 file



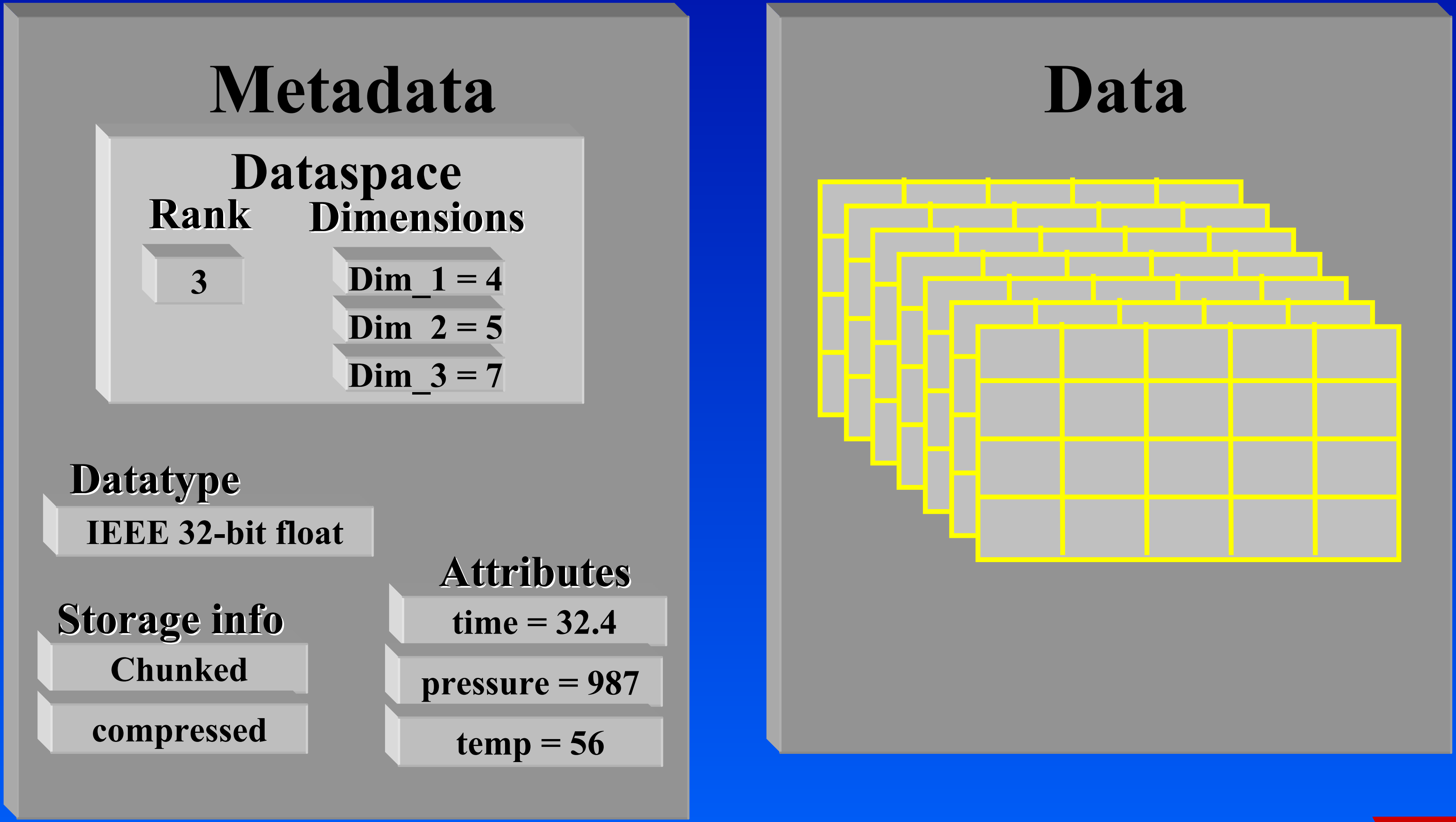
lat	lon	temp
12	23	3.1
15	24	4.2
17	21	3.6

HDF5 file

HDF5 file – container for storing scientific data

- Primary Objects
  - Groups
  - Datasets
- Additional means to organize data
  - Attributes
  - Sharable objects
  - Storage and access properties

Dataset Components



Metadata

Dataspace

Rank

3

Dimensions

Dim_1 = 4

Dim 2 = 5

Dim_3 = 7

Datatype

IEEE 32-bit float

Storage info

Chunked

compressed

Attributes

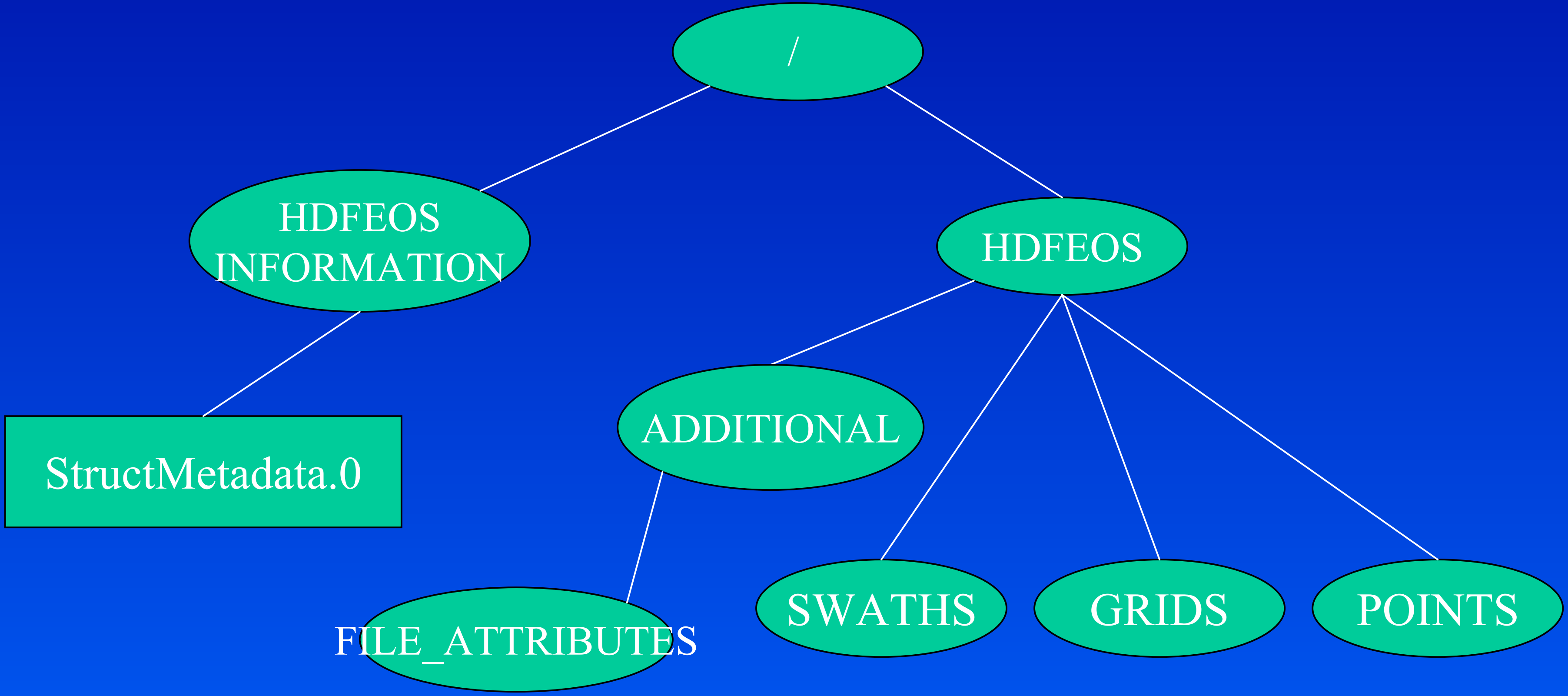
time = 32.4

pressure = 987

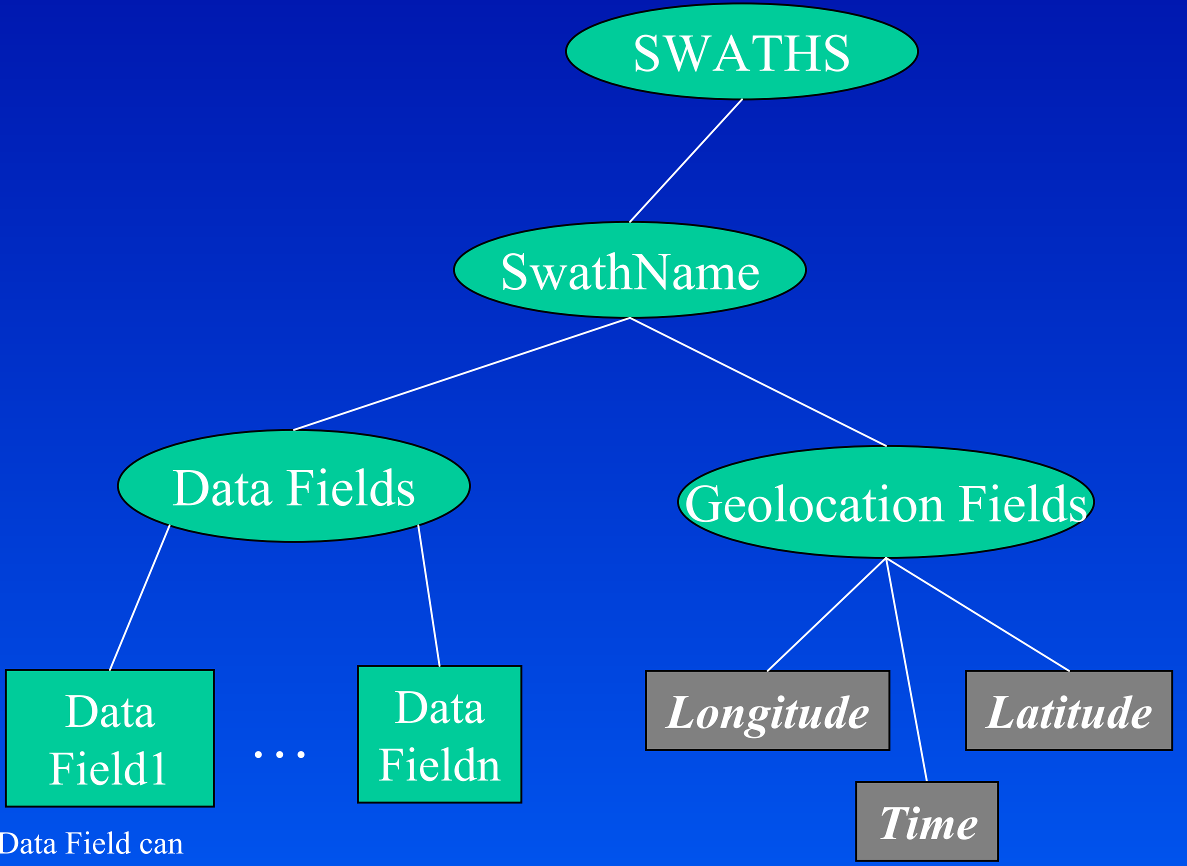
temp = 56

Data

Example
Top Level of HDF-EOS 5 File



Example
Swath Interface for HDF-EOS 5 File



Each Data Field can have Attributes and/or Dimension Scales

Shaded Objects are implemented in a fixed way so the user doesn't have direct access

HDF5 file format

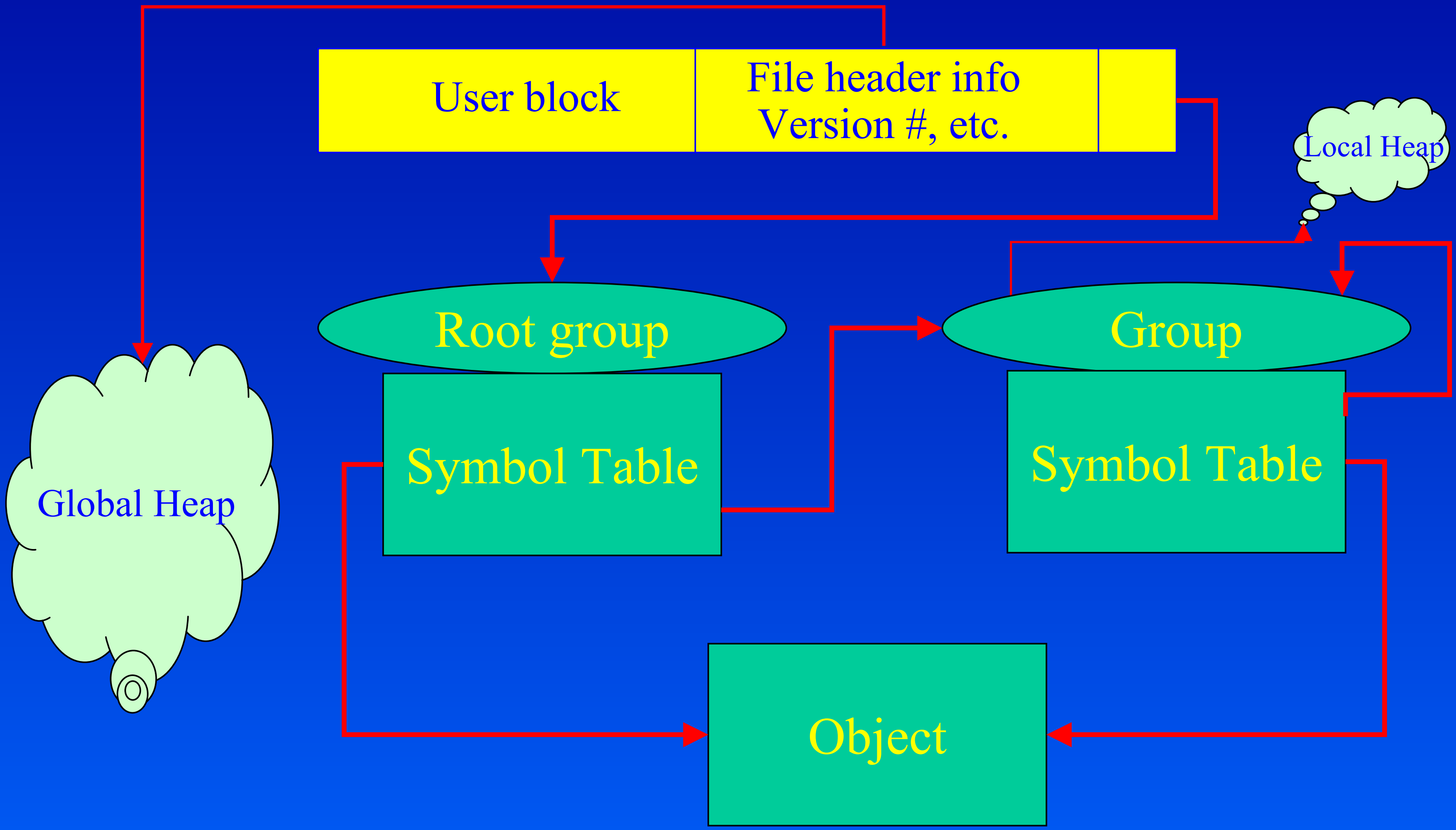
HDF5 File Format

- File Format Specifications
  - http://hdf.ncsa.uiuc.edu/HDF5/doc/H5.format.html
- *Very complex*
- *Need HDF5 Library to write/read data*
- *Describes logical HDF5 file*
- *Does not describe physical storage layout*
- Validation tool h5check

HDF5 File Format

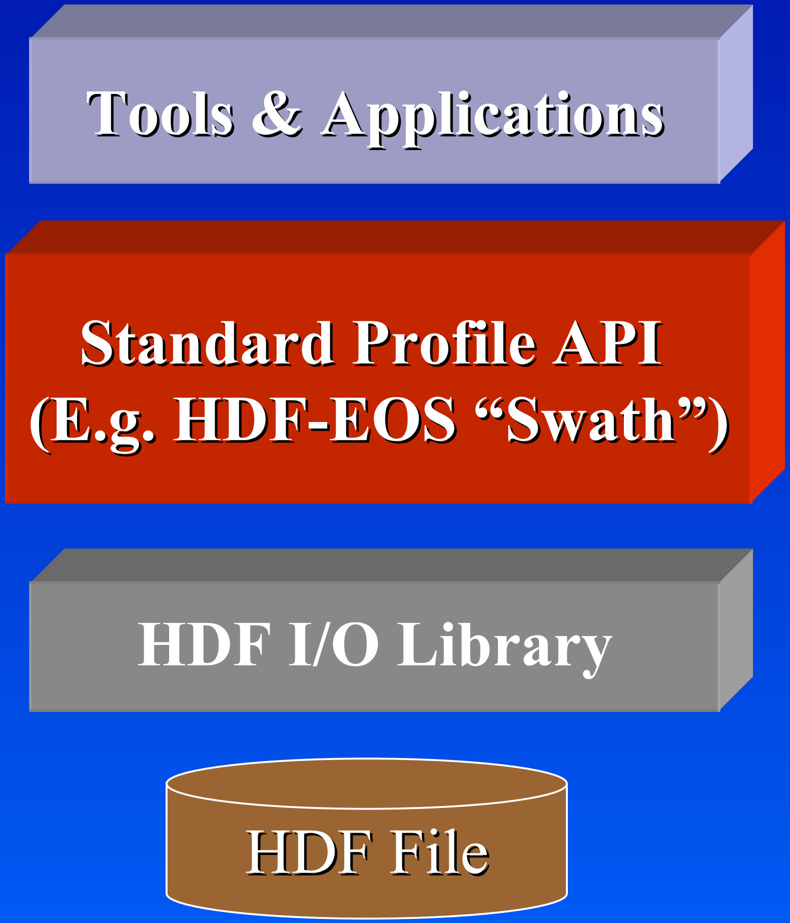
- File Header Block (Super Block)
- B-tree information for storing internal structures
- Symbol Tables
- Local and Global Heaps
- Object headers

HDF5 file structure



HDF5 I/O Library

Supporting a Standard Profile with an API



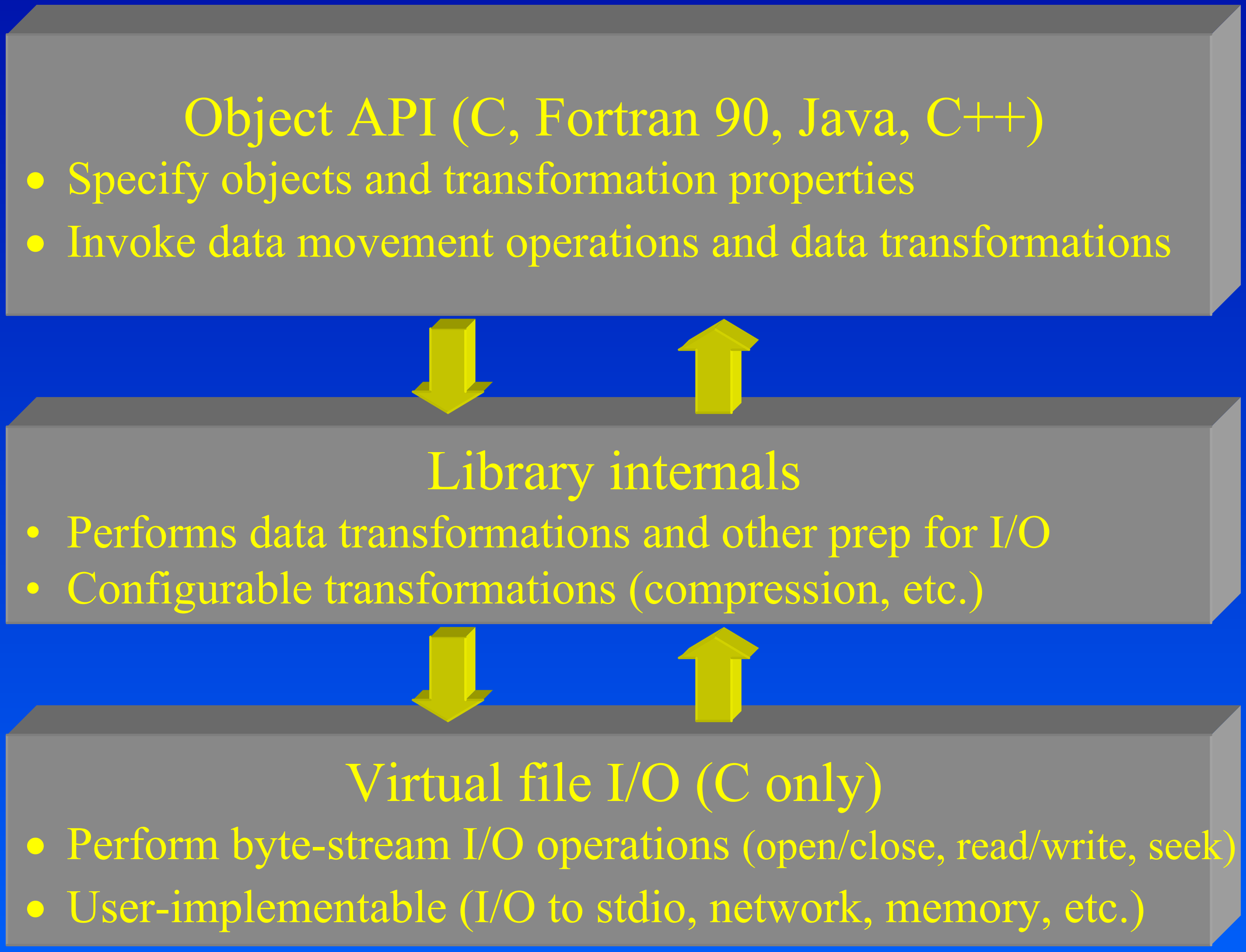
Tools & Applications

Standard Profile API
(E.g. HDF-EOS “Swath”)

HDF I/O Library

HDF File

Structure of HDF5 Library



Object API (C, Fortran 90, Java, C++)

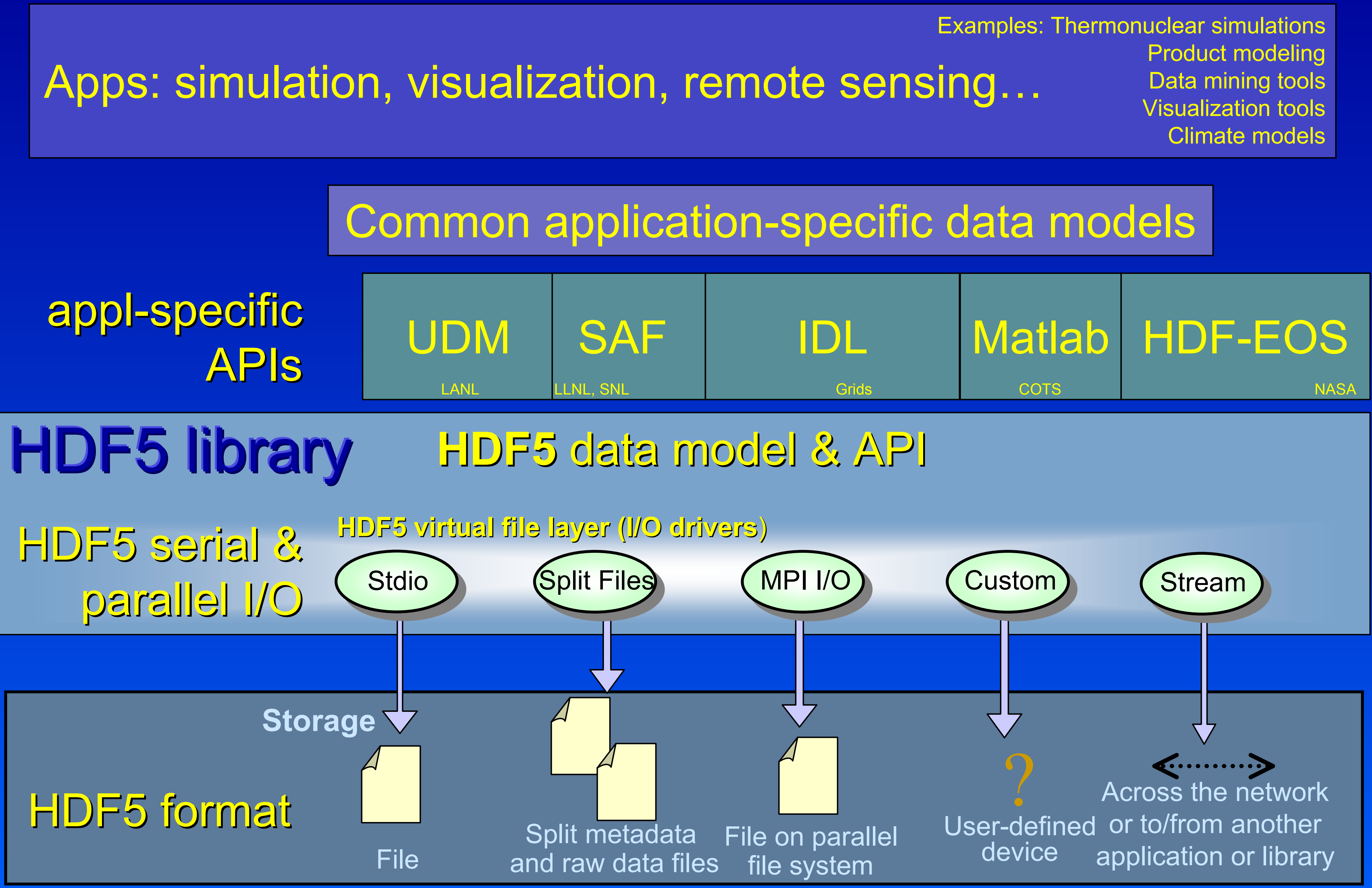
- Specify objects and transformation properties
- Invoke data movement operations and data transformations

Library internals

- Performs data transformations and other prep for I/O
- Configurable transformations (compression, etc.)

Virtual file I/O (C only)

- Perform byte-stream I/O operations (open/close, read/write, seek)
- User-implementable (I/O to stdio, network, memory, etc.)



Apps: simulation, visualization, remote sensing…

Examples: Thermonuclear simulations
Product modeling
Data mining tools
Visualization tools
Climate models

Common application-specific data models

appl-specific APIs

UDM	SAF	IDL	Matlab	HDF-EOS
LANL	LLNL, SNL	Grids	COTS	NASA

HDF5 library

HDF5 data model & API

HDF5 serial & parallel I/O

HDF5 virtual file layer (I/O drivers)

Storage

HDF5 format

File
Split metadata and raw data files
File on parallel file system
?
User-defined device
Across the network or to/from another application or library

HDF5 Existing Implementations and Stability

HDF5 Revisions

- First release in 1998 1.0.0
- Current stable release 1.6.5
  - 1.6.* includes only bug fixes, no new features or changes to file format and data model
- Satisfies NASA TRL9
  - Actual system "mission proven" through successful mission operations (ground or space)

HDF5 Revisions

- Need to address new requirements
  - NetCDF-4
  - Boeing
- Current development branch 1.7.* (future 1.8.0 release)
  - Compact groups
  - UTF-8 encoding
  - Objects creation ordering
  - Datatype conversion between integer and floats
  - Efficient metadata cache implementation

HDF5 Revisions

- Revisions:
  - Bug fixes in a stable branch
  - New features, format changes in development branch
  - Always backward compatible

NASA Technical Readiness Levels

TRL 1 Basic principles observed and reported

TRL 2 Technology concept and/or application formulated

TRL 3 Analytical and experimental critical function and/or characteristic proof-of-concept

TRL 4 Component/subsystem validation in lab environment

TRL 5 System/subsystem/component validation in relevant environment

NASA Technical Readiness Levels

TRL 6 System/subsystem model or prototype demonstration in a relevant end-to-end environment

TRL 7 System prototype demonstration in an operational environment (ground or space)

TRL 8 Actual system completed and "mission qualified" through test and demonstration in an operational environment

TRL 9 Actual system "mission proven" through successful mission operations (ground or space)

“Uses of HDF” from webpage

- Academy of Sciences, Russia
- Accelerators & FEL Physics
- Acoustics
- Advanced Fuel Cycle
- Advanced Parallel Numerical Simulation
- Aeronautic testing
- Aeronautics
- Aerosol Analysis from MISR data
- Aerospace
- Agent Based Modeling

- Air Traffic Control
- Airbourne Laser Project (ABL)
- Aircraft Emissions Database
- Airborne Remote Sensing
- Air-Sea Interaction
- Animal Functional Genomics
- APAC National Facility Scientific Staff
- Applied Mathematics
- Archeology
- Astronomy
- Astrophysics

- Astrophysics/Computation Science
- Atmosphere
- Atmospheric Chemistry
- Atmospheric Physics & Dynamical Meteorology
- Atmospheric radiation
- automotive
- band gap
- Bathymetric database
- Bioeffects of Electromagnetic Fields
- Bioengineering
- Bioinformatics
- biological
- Biomedical Applications
- biomolecular modeling
- biophysics
- biostatistics
- biotech
- Brain Research
- bridge C++/java
- build and install packages for researchers
- cdf
- CEM Simulation
- CERSAT

- CFD
- Chemical Reaction Engineering
- chemistry
- Climate Change
- Climate Modeling
- Climate Physics
- Climatology, Hydrology
- cloud physical
- Comp Accelerator Physics
- computational biology
- computational chemistry
- Comp Electromagnetics
- Computational Grid
- Computational Mathematics
- Computational Physics &
- Computational Astrophysics
- Computational Science/HPC
- Computer Aided Engineering (Durability/Safety)
- computer modeling
- Computers for Nuclear fuel design
- Computer simulation
- Condensed Matter Theory and materials science
- construction of a Beowulf class computer
- Cosmic ray modulation in the heliosphere
- cosmology
- Cplant IO

- data assimilation in meteorology
- Data structures, programming abstraction, and run time support for parallel computation of adaptive and irregular problems
- Data Viewer
- Dawning Cluster File System
- derivation of cloud characteristics
- development of magnetic mass spectrometers
- Diffusion Tensor Magnetic Resonance Imaging
- dislocation dynamics - material modeling
- Displaying WSI's Nexrad Image
- DOD Testing and Evaluation
- Earth and Space Science visualization
- Earth Observation / Remote Sensing
- Earth Observation/Atmospheric Science
- Earth Resources
- Earth Science, Applied Info Sys
- ecology
- EDC/ECS Operations Operator
- education
- Electrical and Computer Engineering
- electromagnetics
- Electronics
- Environmental modeling

- Environmental modeling and soil science
- Environmental monitoring
- Environmental Research
- environmental/meteorological
- Failure Analysis
- Fast searching, sorting and retrieval
- FEL Physics
- final year dissertation
- financial research
- Flight test data analysis
- Fluid Mechanics
- Fluid Dynamics
- Fracture Mechanics
- Fundamental Physics in Space
- Fusion Plasma Physics
- Fusion Science
- Geodetic Science
- Geographic map updating
- Geology
- geology, environment science
- Geophysics
- Geostationary Earth Radiation Budget Sensor

- GERB instrument on MSG satellite
- GIS
- GIS & Remote Sensing
- Globus Project
- GrADS, MicroGrid
- Gravitational physics
- Groundwater heat and solut transport with reactions
- Guinean Research Institute Support
- HIRDLS/AURA
- HPUX Porting and Archive Centre
- Hurricane Research
- Hydrodynamics
- hydraulics
- Hydrology
- ILUMASS - Integrated Land-Use Modeling and Transportation System Simulation
- Image Processing
- Information Technology
- integrated optics

- Interpretation of Aeroelectromagnetic Data
- Land surface modeling (water, carbon)
- Land use and land cover maps in Senegal
- LANL ASCI program
- Laser Plasma experiments
- magnetohydrodynamics
- Marine Biology - Ecology
- marine fishery
- MAT
- material physical
- materials simulation
- Mechanical Engineering; To be used by MIT Photonic Bands package to output the field information
- medical physics
- Meteorology
- Metrology
- Microscopy
- Modeling
- Molecular Biology: Tech Development
- Molecular Level Physiology
- N3C - GAB
- Nano device simulation

- NASA EOS Land Data
- NASA SORCE Mission
- Naval Arch & Marine Engineering
- Neural analog/digital VLSI IC
- Neuroscience research
- Neutron Scattering
- nonlinear optics
- NPOESS Software Development
- Nuclear Engineering
- Numerical fluid dynamics
- numerical general relativity
- numerical relativity
- NYSCEDII
- ocean color
- Ocean Dynamics
- Ocean Physics
- Ocean Remote Sensing
- Oceanography
- Oceanography - surface currents - radar
- Oil & Gas Visualization
- Oil Exploration
- optical information technology
- Optical Integrated Circuit Design

- Optics
- Optoelectronics
- Ozone Monitoring Instrument
- Parallel Programming
- PDE
- Performance Modeling and Evaluation
- Petroleum Engineering
- petrology
- photonic band gap studies
- Photonic Crystals, INFM PRA
- Photonics
- Physical Oceanography
- physics
- Physics (Medical Physics)
- physics / numerical simulation
- Physics, Optics
- Physics, Photonic band gaps
- Physics/Materials Science
- Physics, Optics, Metrology
- Plasma / Electromagnetics
- Plasma Physics

- Plasma Physics - Nuclear Fusion
- Polymer physics
- Post-fire erosion analysis
- Processing of (weather) radar images
- Process-understanding arsenic groundwater contamination in West Bengal/Bangladesh
- protein crystallography, molecular modeling
- Protostellar accretion discs
- Quantitative Precipitation Forecast
- Quantum Chemistry
- Radar
- Radar and Satellite data processing
- Radar Science
- remote science
- Remote Sensing
- Remote sensing, satellite meteorology
- Research
- Robotics
- SAF
- SAMRAI
- SAR processing
- SATC

- satellite
- Satellite Climate Monitoring
- satellite Image
- satellite image viewer
- Satellite meteorology
- Satellite oceanography
- satellite remote sensing
- Satellite/weather radar remote sensing
- Scanning Near Field Optical Microscopy
- SciDAC
- science
- scientific graph and data analysis
- Sea Ice
- seafloor modeling
- Seismology
- Semiconductor Process Simulation
- Ship hydrodynamics
- Signal Processing
- signal processing
- SNOW melting
- Software Engineering, Distributed Systems
- Solid State Physics
- Southern Ocean predator ecology

- Space Geodesy
- Space Modeling
- Space Physics
- Space Plasma Physics
- Space Sciences
- Statistics
- Student just looking for a way to view hdf files
- supercomputing
- Support to geophysical applications
- Surface water flow and sediment transport
- SWX: Space Weather Explorer
- System support/integration/flight test
- Systems Analysis
- telecommunications
- theoretical chemistry
- To retrieve SST from MODIS data in Indian Ocean
- trabajo académicamente dirigido
- TRMM
- Underwater acoustics
- vision research, neuroscience
- Visualization
- Visualization and Analysis Software
- Volcano Simulation project

- Volcanology
- Water Modeling
- Water Resources Management
- Weather Data Display
- X-ray physics